Linear regression Advertising data

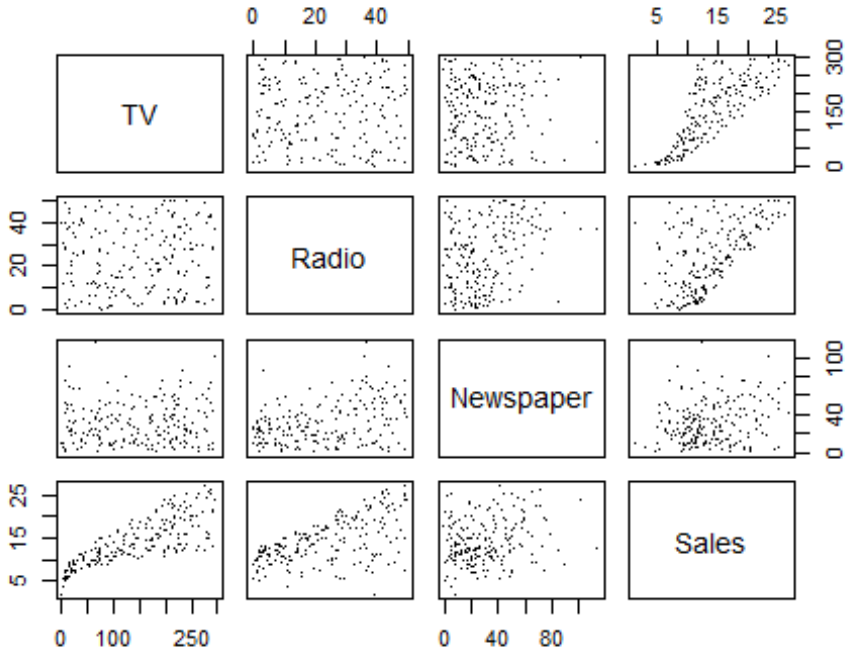
Linear Regression

```
library(MASS)

library(ISLR)

## Warning: package 'ISLR' was built under R version 3.4.3

Advertising<-read.csv("Advertising.csv")
wd <- getwd(); file <- paste(wd,"Advertising.csv",sep="/")
Advertising<-read.csv(file, head=TRUE)[,-1]
```

Advertising data sales (in thousands of units) for a particular product advertising budgets (in thousands of dollars) for TV, radio, and newspaper media

On the basis of this data,suggest a marketing plan for next year that will result in high product sales.

```
head(Advertising)

##      TV Radio Newspaper Sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
## 6   8.7  48.9      75.0   7.2

summary(Advertising)

##        TV              Radio          Newspaper          Sales      
##  Min.   :  0.70   Min.   : 0.000   Min.   :  0.30   Min.   : 1.60  
##  1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75   1st Qu.:10.38  
##  Median :149.75   Median :22.900   Median : 25.75   Median :12.90  
##  Mean   :147.04   Mean   :23.264   Mean   : 30.55   Mean   :14.02  
##  3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10   3rd Qu.:17.40  
##  Max.   :296.40   Max.   :49.600   Max.   :114.00   Max.   :27.00  

pairs(Advertising, pch=".")
```



1. Is there a relationship between advertising sales and budget?

A multiple regression model of sales onto TV, radio, and newspaper

```
ad.lm <- lm(Sales~., data=Advertising)
summary(ad.lm)
```

```
##
## Call:
## lm(formula = Sales ~ ., data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
```

```
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Test hypothesis H0 : beat_TV = beta_radio = beta_newspaper = 0. F-test indicates clear evidence of a relationship between advertising and sales

2. How strong is the relationship?

Two measures of model accuracy:RSE and R^2 RSE: Residual (y-yhat) standard error

```
rse=summary(ad.lm)$sigma

#RSE= 1.686
mean(Advertising$Sales)

## [1] 14.0225

rse/mean(Advertising$Sales)

## [1] 0.1202004

# noisy/signal=.12
#percentage error wrt mean is roughly 12 %.
#

rsq=summary(ad.lm)$r.sq
rsq #0.8972106

## [1] 0.8972106
```

The predictors explain almost 90 % of the variance in sales

```
# rsq is calculated by the following formuala
yhat=ad.lm$fitted.values #predicted
y=Advertising$Sales #observed
rsq=1-sum((y-yhat)^2)/sum((y-mean(y))^2) #orginal formula


#Other way to get R2
var(yhat)/var(y) #other formula

## [1] 0.8972106

1-sum((y-yhat)^2)/sum((y-mean(y))^2) #orginal formula

## [1] 0.8972106

cor(yhat,y)^2 #alternate formula

## [1] 0.8972106
```

They are equal for linear regression model.

3. Which media contribute to sales?

```
Coef1=summary(ad.lm)$coefficients #Coefficient matrix
Coef1
```

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
## TV           0.045764645 0.001394897 32.8086244 1.509960e-81
## Radio        0.188530017 0.008611234 21.8934961 1.505339e-54
## Newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
```

Examining the p-values associated with each predictor's t-statistic, the p-values for TV and radio are low, but the p-value for newspaper is not. This suggests that only TV and radio are related to sales.

4. How large is the effect of each medium on sales?

```
lolim=Coef1[,1] - 1.96*Coef1[,2]
uplim=Coef1[,1] + 1.96*Coef1[,2]
cbind(lolim,uplim)
```

```
##                   lolim      uplim
## (Intercept)  2.32754923 3.55022951
## TV           0.04303065 0.04849864
## Radio        0.17165200 0.20540804
## Newspaper   -0.01254467 0.01046969
```

```
confint(ad.lm)
```

```
##                   2.5 %     97.5 %
## (Intercept)  2.32376228 3.55401646
## TV           0.04301371 0.04851558
## Radio        0.17154745 0.20551259
## Newspaper   -0.01261595 0.01054097
```

The confidence intervals for TV and radio are narrow and far from zero, providing evidence that these media are related to sales. But the interval for newspaper includes zero, indicating that the variable is not statistically significant given the values of TV and radio.

Could collinearity be the reason that the confidence interval associated with newspaper is so wide? Variation Inflation factor (vif) measures collinearity: vif(betaj hat)=1/(1-R^2) where R^2 is from the regression of Xj all the other predictors. Rule of thumb: vif>5 or 10 indicates problematic collinearity.

```
require(car)
```

```
## Loading required package: car
```

```
vif(ad.lm)
```

```
##        TV     Radio Newspaper
##  1.004611  1.144952  1.145187
```

The VIF scores are 1.005, 1.145, and 1.145 for TV, radio, and newspaper, suggesting no evidence of collinearity.

5. How accurately can we predict future sales?

For individual response, a prediction interval is used, and for the average response, f(X) for the average response f(X)

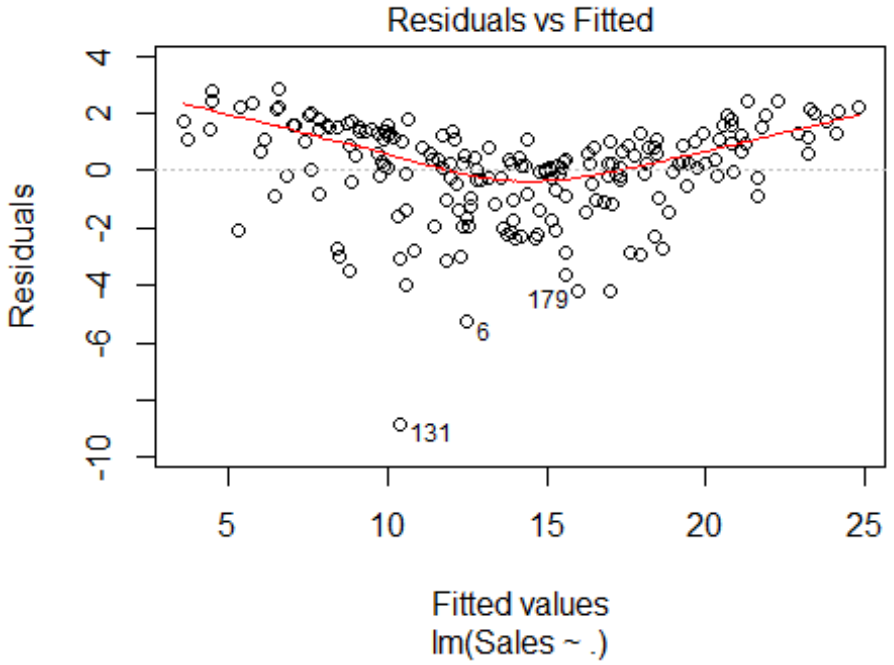
```
#for the  average  response f(X)
predict(ad.lm, newdata=data.frame(TV=149,Radio=22,Newspaper=25),
        interval="confidence")

##        fit      lwr      upr
## 1 13.87954 13.63678 14.12231

#for individual  response
predict(ad.lm, newdata=data.frame(TV=149,Radio=22,Newspaper=25),
        interval="prediction")

##        fit      lwr      upr
## 1 13.87954 10.54663 17.21246
```

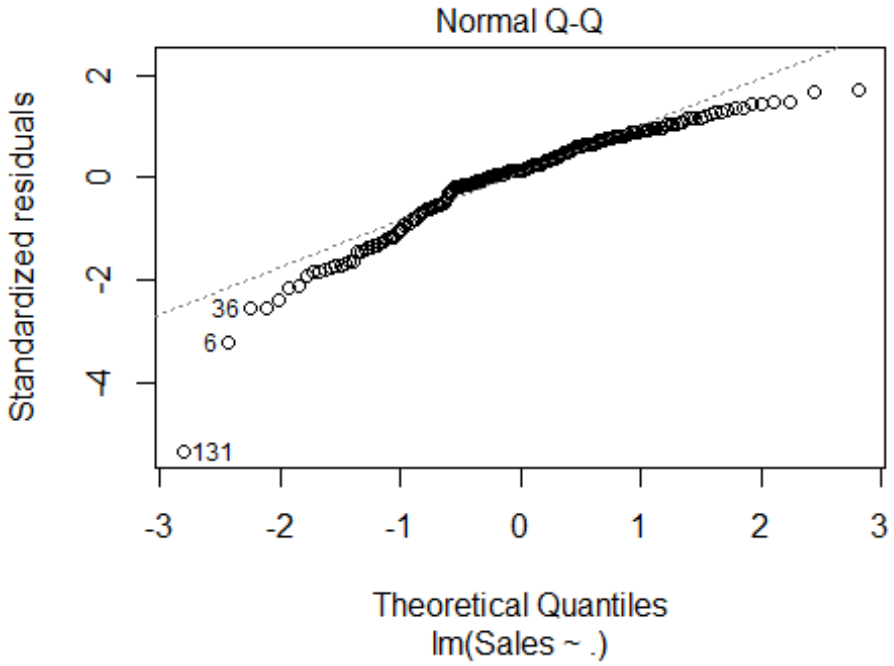
Prediction intervals are always wider than confidence intervals because they account for the uncertainty associated with epsilon e, the irreducible error.

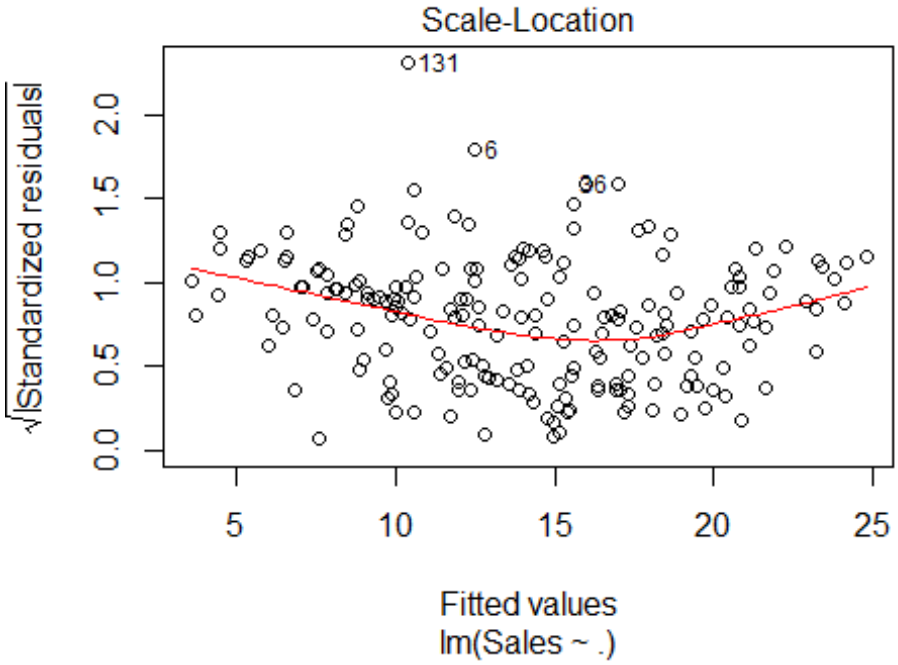
6. Is the relationship linear?

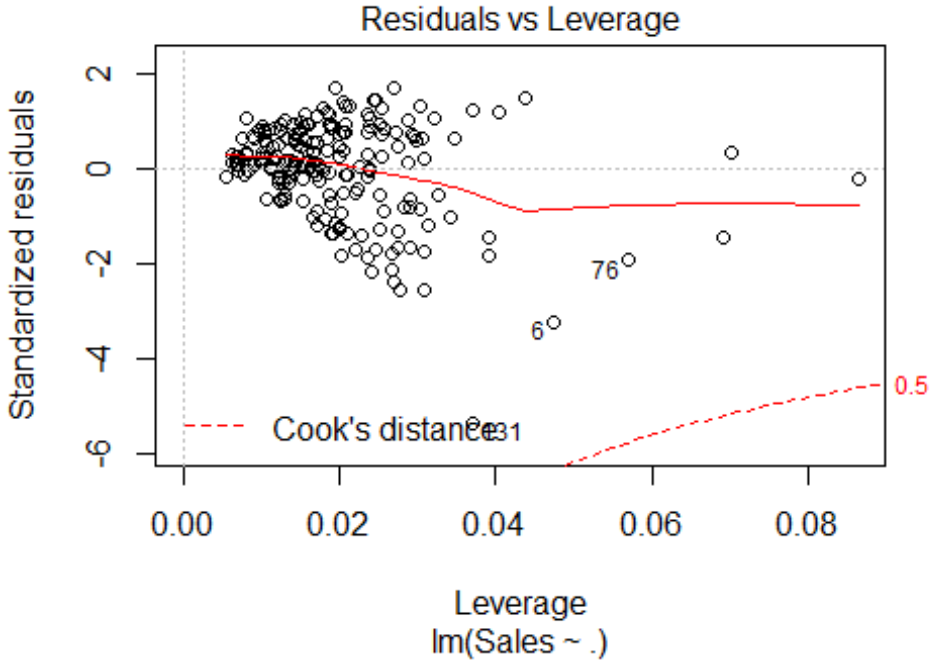
If the relationships are linear, then the residual plots should display no pattern.

```
plot(ad.lm) #diagnostic plot
```


Residuals vs Fitted
Residuals
Fitted values
lm(Sales ~ .)
179
6
131


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Sales ~ .)
36
6
131





7. Is there synergy among the advertising media?

non-additive relationships model

```
ad.lm2 <- lm(Sales~.^2, data=Advertising)
summary(ad.lm2)

##
## Call:
## lm(formula = Sales ~ .^2, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9239 -0.3954  0.1873  0.5976  1.5267
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.460e+00  3.176e-01  20.342   <2e-16 ***
## TV               2.033e-02  1.609e-03  12.633   <2e-16 ***
## Radio            2.293e-02  1.141e-02   2.009   0.0460 *
## Newspaper        1.703e-02  1.007e-02   1.691   0.0924 .
## TV:Radio         1.139e-03  5.716e-05  19.930   <2e-16 ***
## TV:Newspaper    -7.971e-05  3.579e-05  -2.227   0.0271 *
## Radio:Newspaper -1.096e-04  2.363e-04  -0.464   0.6433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9383 on 193 degrees of freedom
## Multiple R-squared:  0.9686, Adjusted R-squared:  0.9677
## F-statistic: 993.3 on 6 and 193 DF,  p-value: < 2.2e-16
```

The Advertising data may not be additive.

```
summary(ad.lm2)$r.sq;summary(ad.lm)$r.sq

## [1] 0.9686311

## [1] 0.8972106
```

Including an interaction term in the model results in a substantial increase in R2, from around 90 % to almost 97 %.

Non-linear Transformations of the Predictors

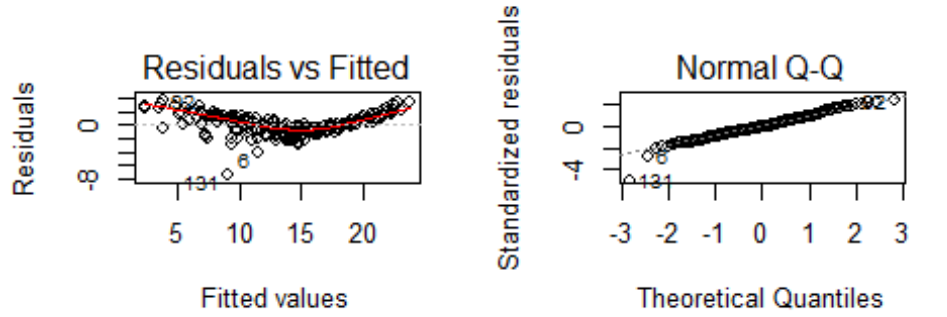
```
ad.lm3 <- lm(Sales~.+I(TV^2), data=Advertising)
summary(ad.lm3)

##
## Call:
## lm(formula = Sales ~ . + I(TV^2), data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3583 -0.8701 -0.0484  0.9562  3.5604
##
## Coefficients:
```

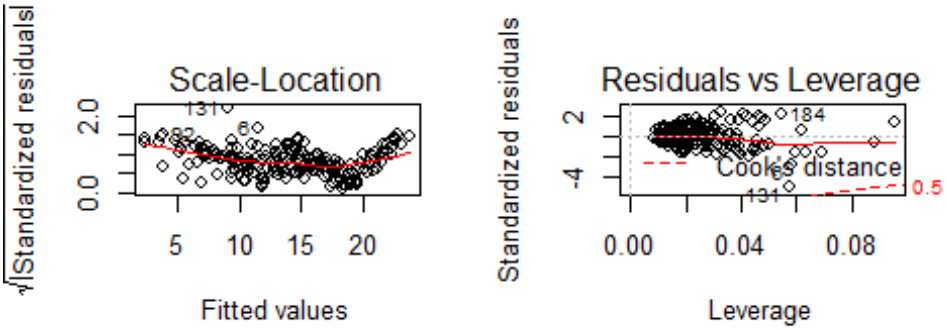
```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.270e+00  3.745e-01   3.392  0.00084 ***
## TV           7.847e-02  5.001e-03  15.690  < 2e-16 ***
## Radio        1.926e-01  7.794e-03  24.706  < 2e-16 ***
## Newspaper    8.906e-04  5.306e-03   0.168  0.86688
## I(TV^2)     -1.137e-04  1.683e-05  -6.757 1.59e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.521 on 195 degrees of freedom
## Multiple R-squared:  0.9167, Adjusted R-squared:  0.915
## F-statistic: 536.6 on 4 and 195 DF,  p-value: < 2.2e-16
```

```
anova(ad.lm,ad.lm3)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ TV + Radio + Newspaper
## Model 2: Sales ~ TV + Radio + Newspaper + I(TV^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    196 556.83
## 2    195 451.19  1    105.64 45.656 1.587e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Indicates the non-linear effect of TV

par(mfrow=c(2,2))
plot(ad.lm3)
```






```
ad.lm4 <- lm(Sales~.+poly(TV,3), data=Advertising)
summary(ad.lm4)

##
## Call:
## lm(formula = Sales ~ . + poly(TV, 3), data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1989 -0.8342 -0.0653  0.7703  3.7311
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.753e+00  2.657e-01  10.362  < 2e-16 ***
## TV            4.568e-02  1.184e-03  38.574  < 2e-16 ***
## Radio         1.961e-01  7.365e-03  26.629  < 2e-16 ***
## Newspaper    -3.371e-04  4.997e-03  -0.067    0.946
## poly(TV, 3)1         NA         NA      NA       NA
## poly(TV, 3)2 -1.039e+01  1.441e+00  -7.212 1.20e-11 ***
## poly(TV, 3)3  7.378e+00  1.438e+00   5.133 6.91e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.431 on 194 degrees of freedom
## Multiple R-squared:  0.9267, Adjusted R-squared:  0.9248
## F-statistic: 490.3 on 5 and 194 DF,  p-value: < 2.2e-16
```

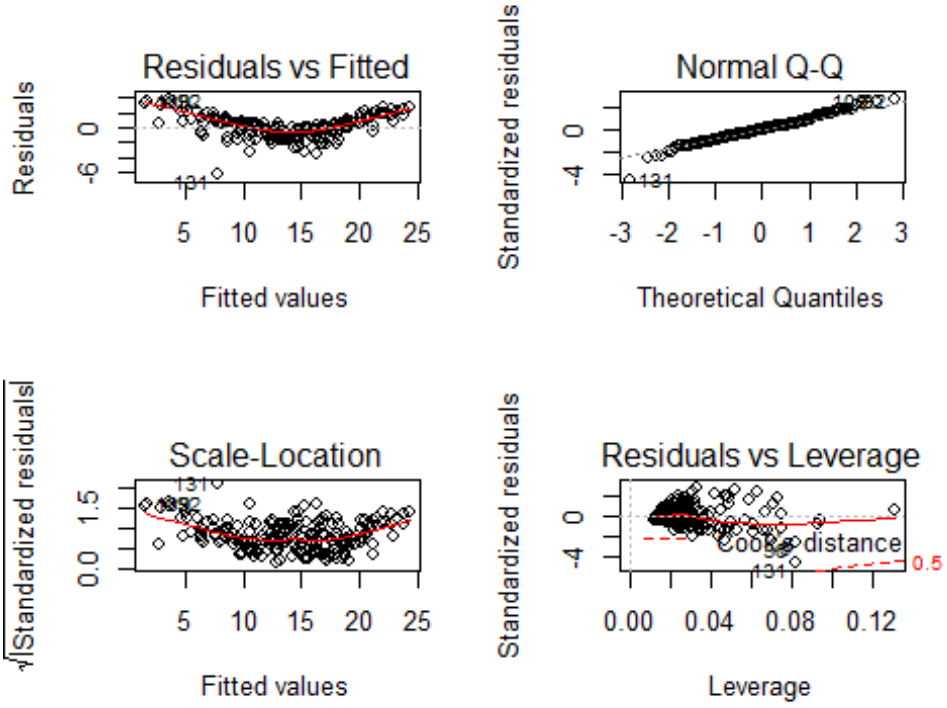
```r
anova(ad.lm,ad.lm4)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ TV + Radio + Newspaper
## Model 2: Sales ~ TV + Radio + Newspaper + poly(TV, 3)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    196 556.83
## 2    194 397.25  2    159.58 38.967 5.942e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

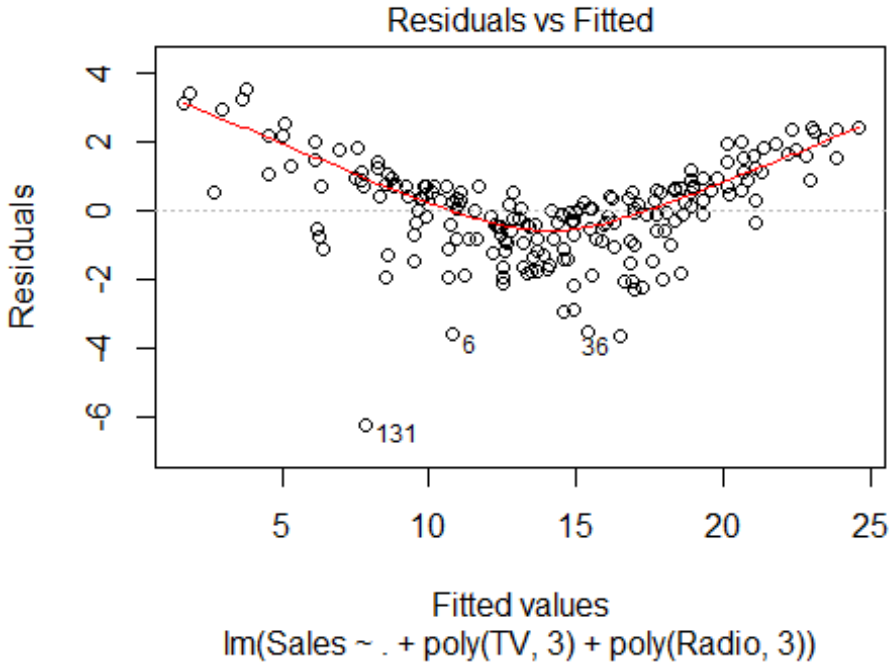
```r
anova(ad.lm3,ad.lm4)
```

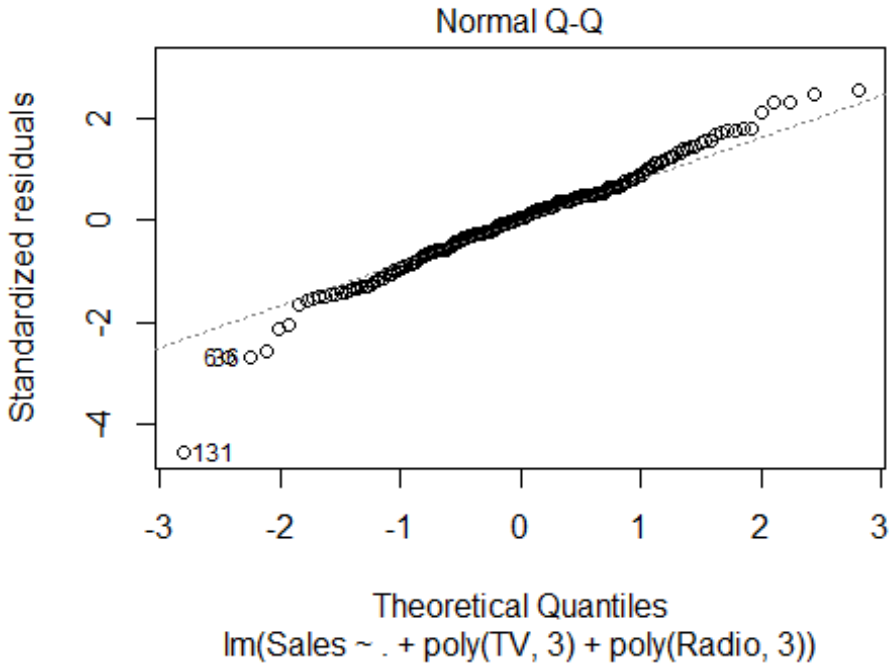
```
## Analysis of Variance Table
##
## Model 1: Sales ~ TV + Radio + Newspaper + I(TV^2)
## Model 2: Sales ~ TV + Radio + Newspaper + poly(TV, 3)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    195 451.19
## 2    194 397.25  1    53.942 26.343 6.915e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

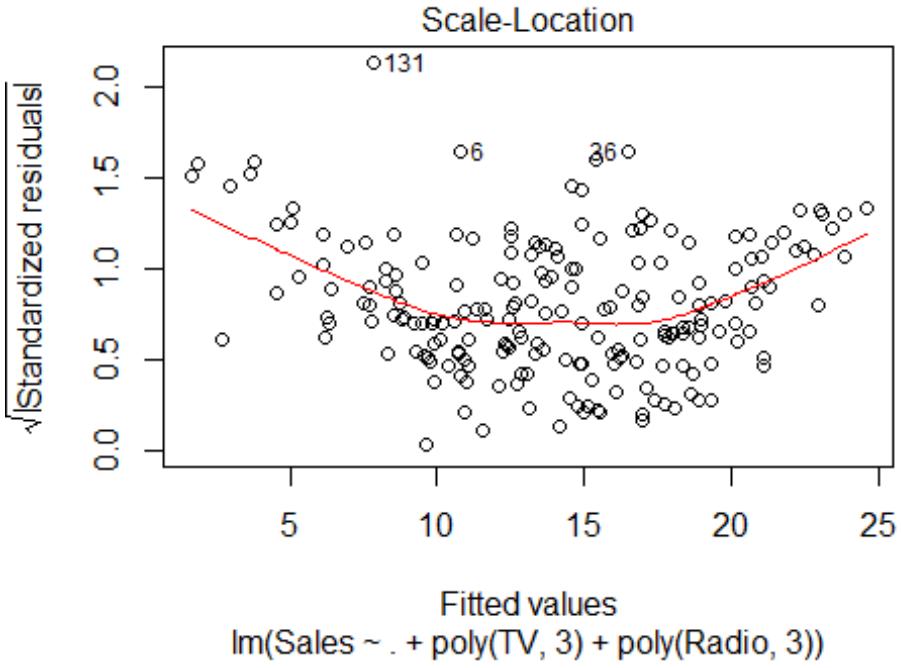
```r
par(mfrow=c(2,2))
plot(ad.lm4)
```

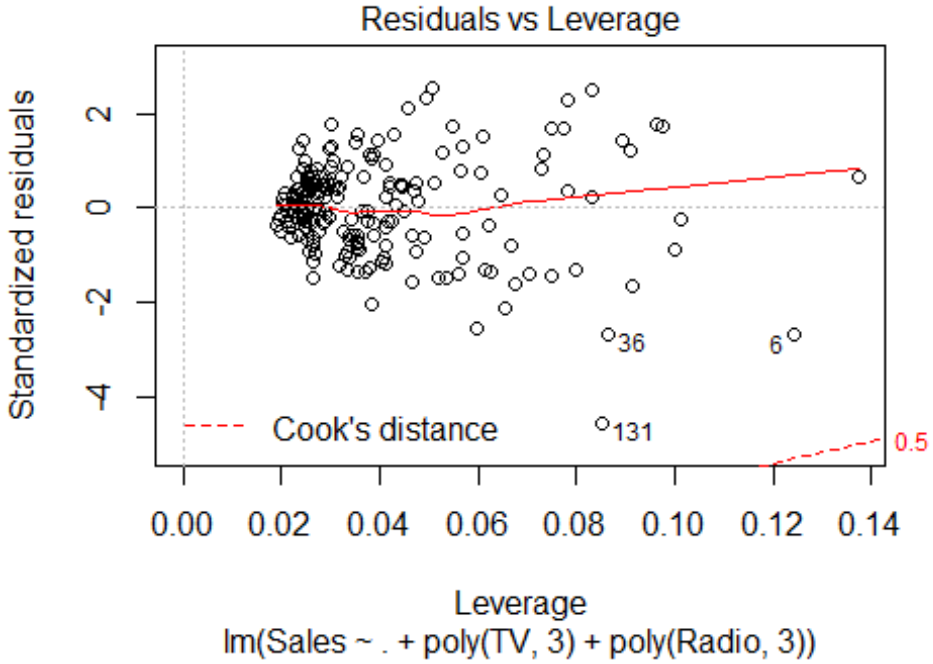


```
ad.lm5 <- lm(Sales~.+poly(TV,3)+poly(Radio,3), data=Advertising)
plot(ad.lm5)
```










```
anova(ad.lm4,ad.lm5)

## Analysis of Variance Table
##
```

```
## Model 1: Sales ~ TV + Radio + Newspaper + poly(TV, 3)
## Model 2: Sales ~ TV + Radio + Newspaper + poly(TV, 3) + poly(Radio, 3)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    194 397.25
## 2    192 394.25  2    2.9929 0.7288 0.4838
```